Web Scraping and Social Media Scraping Project

Krzysztof Kalisiak & Maciej Zdanowicz
409879 411968

The aim of this project is to gather data regarding the results of matches in the highest level of men's volleyball – PlusLiga. As volleyball is becoming more and more popular, we believe that statistics regarding polish teams might be of interest for plethora of people. Website used is the *https://www.plusliga.pl/* - the official website of PlusLiga league. This website contains many different information regarding the volleyball news, tickets, performance of individual players, teams and matches, as well as schedule of upcoming matches. Out of the wide range of information we have decided to gather data concerning the results of individual matches for every team in all possible season. Apart from the score alone, each match statistics include: number of points gathered, data regarding the quality of serving, reception, attack and block.

As we have decided to gather data for individual teams, each scraper starts at the *https://www.plusliga.pl/statsTeams.html* website. Then proceeds to visit one of the team's website displayed on the page, chooses one of the possible seasons with available statistics and downloads data. This procedure is repeated for all seasons, and for all possible teams. Even though the three provided scrapers are based on different libraries and their functionalities are based on different principles, the idea for each of them is similar – find the website associated with each team, find particular season, find the table with results, download the data, and repeat.

The final output, due to its size – we have obtained data for 29 teams for all matches available, for some teams that means over 400 games - is stored in an sql database. All three scrapers gather exactly the same information. Out of three scrapers used scrapy was definitely the fastest one (even on windows), while both selenium and beautiful soup were considerably slower.

The work on the project was divided as follows:
-Beautiful Soup – Maciej Zdanowicz
-Scrapy – Krzysztof Kalisiak
-Selenium – Krzysztof Kalisiak & Maciej Zdanowicz
-Description file and Read.me - Krzysztof Kalisiak & Maciej Zdanowicz

Below there are some sample graphs created from the subsamples of gathered data. First graph presents the number of points gathered by Cuprum Lubin in first 15 matches respectively. Second graph represents the pie chart representing the frequency of matches played by Cuprum Lubin against 10 teams from Plusliga.